GÖTEBORGS UNIVERSITET

What makes a good language test in EFL?

Jani Lemmetti
Programme Teacher Education

Uppsats/Examensarbete: 15 hp
Kurs: LGEN1G
Nivå: Grundnivå
Termin/år: HT2014
Handledare: Anne Dragemark Oscarsson
Examinator: Monika Mondor
Kod: HT14-1160-009-LGEN1G

Key words: Language testing, validity, reliability, washback

Abstract

Language testing is a fundamental part of learning and teaching in school today, and has been throughout history even though views on language testing have changed. This paper reports on what research says regarding the various components that are needed when constructing and using a language test. The findings points towards the importance of validity, reliability, and washback and the fact that these issues should be addressed with high consideration in order for a test to have a positive effect. We can see that evidence points to the fact that when tests are used, they have to measure what they are supposed to measure and that the evidence in validity is crucial. Furthermore, the terms test-retest and parallel tests were emphasized when discussing the reliability concept even though those methods have problems. Moreover, when the concept of washback was examined, it was clear that it is a powerful tool for both language learners and teachers. The literature suggested that the focus should be on impact and not processes. Finally, the presented criticism towards certain language tests showed that the tests were not used to assess language proficiency, and had both reliability and validity issues. As it seems, most classroom tests are neither very reliable nor possibly valid because teachers are not able to construct proper tests with all these features. The results of this review seem to indicate that there is a lack of research regarding on how this gap could be closed and therefore deserves more attention.

Contents

Contents **1**

1 Introduction **2**

2 Background **2**

The paradigm shift 3

The western world and formative and summative assessment 3

1970s Sweden and technology development 3

Present time 4

3 Research question **5**

4 Method **5**

5 Validity, reliability and washback **5**

5.1 The validity of testing 5

5.1.1 Validity as a Unitary concept 6

5.1.2 Validity: testing the test 6

5.1.3 Evidence in validity 11

5.2 Reliability 12

5.2.1 Different types of reliability 13

5.3 Washback 14

5.3.1 Impact on test takers 16

5.3.2 Impact on teachers 16

5.3.3 Impact on society and education system 17

6 Conclusion and discussion **17**

Reference list **20**

1 Introduction

Testing can be conducted in various ways, and with different approaches such as written exams, essays, oral exams, conversations between teacher and student or group conversations. But what does testing really measure? And what different aspects do we have to consider when conducting formal or informal assessment and testing? In school students are assessed in many different ways. Formal and informal assessment is an ongoing everyday process, which consist of classroom observations, oral questions, or different kinds of written tests. A written test is a classic example of formal assessment where the student is aware of the fact that he or she is being tested for a reason. Giving the students tests could be both positive and negative. McNamara (2000), describes testing as a universal feature of social life and that testing for special purposes or to establish identity has become an accepted part of many fields, such as sport (drugs testing), the law (DNA tests), medicine (blood tests), and other fields. What is true of testing in general is true also in language testing. McNamara also states that in many cases taking a test causes reactions and most likely, the reactions will be of a negative kind. For many, the view of language testing is associated with a classroom, a traditional paper and pen test, and a race against the limitation of time that is given. However, there is far more to language testing than this. The main focus and research question of this paper will be (as the title suggests) on the different aspects that need to be considered when designing and using a language test. In addition to that, this paper will also include some background on how the view of testing has changed throughout the years. The concepts that will be discussed are *validity*, *reliability*, and *washback.* These terms will need to be addressed in any account of testing. Also, the Swedish perspective and the relevance is something that will be further developed in the discussion.

2 Background

This paper will not explore the history of testing to any great extent. However, it is important to know that the view of testing has shifted focus from fifty years ago up until present time.

The paradigm shift

Many researchers talk of the paradigm shift when it comes to testing, from psychometrics (questionnaires, tests, raters' judgements, and personality tests) to a broader model of assessment in education, from a testing and examination culture to an assessment culture. Gipps (1990) explains this by saying that there is now a wider range of assessment practices in use than there was twenty-five years ago. Examples of these assessment practices are: teacher assessment, standard tasks, coursework, records of achievement, and practical and oral assessment. Written examinations and standardized tests are more common to use now than before. Testing and assessment have a higher profile now, and is required in order to achieve various purposes: support teaching and learning, provide information about the students, teachers and schools, be selective and certifying, and drive curriculum and teaching. These new forms and purposes mentioned by Gipps mean that the major traditional assessment theory, the psychometric model is not adequate anymore, henceforth the paradigm shift.

The western world and formative and summative assessment

As can be seen historically in Dragemark Oscarson (2009), language education and language assessment has followed the same general trend and pattern in the western world. Language education and language assessment are highly influenced by research in fields such as linguistics, socio-linguistics, psychology, and sociology.

Theories and beliefs about general learning are closely related to predominant testing and assessment practices. Further, Dragemark Oscarson discusses the terms *formative* assessment and *summative* assessment. Formative assessment is often referred to as assessment *for* learning and is mainly used to provide information for the student regarding where he/she is in the learning process and how to move further. The goal is to improve the learning whereas summative assessment is more assessment *of* learning. In this case to sum up end results of achievement, and a way of doing that is to use different types of tests.

1970s Sweden and technology development

At the beginning of the 1970s, the view on education became more radical in the western world. This had an impact on the Swedish school system as well. In Sweden, the SIA (Skolans Inre Arbete) reform was implemented, which stressed that students with different social

backgrounds should meet in the same classroom and that education would have a stronger connection to society and to a greater extent make the students' everyday experiences a starting point for schoolwork.
However, assessment and testing did not go through any major changes during this period of time, even though it was common to experiment with education regarding society and everyday experiences. As a result the tests reproductive characteristics were strengthened because the so called education technology had a huge impact in schools. Through this technology with its pre-programmed teaching aids, the tests in addition to just having been used in a selective way, could now be used as pedagogical tools. For instance, it became common to use course books that had diagnostic tests and key included in its content (at the end of each chapter) so that the students could assess their progress and decide for themselves if they were ready to move on or if they needed more practice. This form of individualized learning gave the teachers more allotted time and increased the students' motivation (Korp, 2003).

Present time

Today in the 21st century, changes in the economical structure have effected the role of the educational system in relationship to both society and all individuals who enroll in education. By that, the question was raised: what consequences could this change have when it comes to the functionality of assessment in society, classroom, and the individual? The need to find new ways to motivate especially low-achieving students became more urgent.
Otherwise they might end up in unemployment and social marginalization. Against this background one might say that it is logical that current discussions regarding testing are not so much about selection and competition, but more towards “deep learning”, self understanding, and individualized learning (Korp, 2003).

Another perspective that has been prominent in the discussions and research about assessment puts the “effectiveness” of schools as a dominant factor. Effectiveness in this case means that the students' grades and test results are used as a measurement to indicate how effective that particular school is. The American researchers Madaus and O´Dwyer (in Korp,2003) describes how the outcome of this type of testing in schools in USA and England are used as a basis to make decisions about for example the teachers salaries and whether to close down schools which do not “measure up” (Korp,2003).

3 Research question

This literature review focuses on the different aspects we need to consider when constructing a language test. What does the literature say about the various components such as validity, reliability, and washback?

4 Method

This literature review is based on research which has been collected through various articles online from the library of the University of Gothenburg (GUNDA) and the European Association for Language Testing and Assessment (EALTA) website. One major database has been used in the search for relevant literature. ERIC (Educational Resources Information Center) which is sponsored by the U.S. Department of Education to provide extensive access to educational-related literature. The search keywords have been *language testing, validity, reliability,* and *washback.* In addition to that, various written books regarding language testing has also been examined and used to present the relevant findings.

5 Validity, reliability and washback

The following chapters will now present the research findings within the different fields that need to be considered when designing a language test. First validity will be accounted for. Secondly, reliability is presented, and finally washback. Not only are the different concepts described, but the literature review also presents various critique against certain language tests.

5.1 The validity of testing

The most traditional definition of validity is the extent to which a test really measures what it is supposed to measure. If it does not meet that purpose then testing could be useless or misleading. Four types of validity are emphasized when looking at early readings: predictive validity, content validity, construct validity, concurrent validity (Gipps, 1994).

- *Predictive validity* relates to whether the test predicts future performances accurately or well.

- *Content validity* covers the more appropriate and necessary content which is necessary for a good performance.
- *Construct validity* relates to whether the test is actually adequate to what is being assessed.
- *Concurrent validity* is whether a test correlates with or gives nearly the same result as another similar test of the same skill.

However, emphasis on these different types of validity has led to a situation where evidence might point to only one or perhaps two of these various validity types when developing tests (Gipps, 1994).

5.1.1 Validity as a unitary concept

According to Gipps (1994), recent literature on validity has expressed that first of all validity should be addressed as a unitary concept with construct as the unifying theme. Secondly, the responsibility for valid test use is now placed on the test user (the teacher) and not the developer (then it has to have construct validity). Finally, validity rather than technical reliability is the emphasis on developing performance assessment, which is a reversed situation with for example standardized tests. Messick (in Gipps, 1994) describes the testing profession´s move towards recognizing

> *validity as a unitary concept, in the sense that score meaning as embodied in construct validity underlies all score-based inferences. But for a fully unified view of validity, it must also be recognized that the appropriateness, meaningfulness and usefulness of score-based inferences depend as well on the social consequences of the testing. Therefore, social values cannot be ignored in considerations of validity* (Gipps, 1994, p.59).

5.1.2 Validity: testing the test

Testing is a matter of using data to establish evidence of learning. According to McNamara (2000), evidence does not only occur in the natural state, but also in an abstract inference and a matter of judgement. He draws parallels between testing and legal procedures (by using the OJ Simpson trial as an example) because the question he raises is who makes the judgement and how can we decide how valid the evidence is. Furthermore, McNamara states that these two

stages are mirrored in language test development and validation and that the purpose of validation in language testing based on test performances is that it can be defendable and fair. Test validation is about the logic of the test, and especially the design and intention. Moreover, it involves looking at empirical evidence emerging from data from test trials or operational administrations. It might be unfair and unjust if there are no validation procedures available. Considering what might be at stake, these procedures must be addressed with importance. Hughes (2003) also claims that empirical evidence is needed and that it is not enough to state that a test has construct validity without the proper empirical evidence. Furthermore, he states that evidence, and especially the subordinate forms of validity *content validity* and *criterion-related validity* are essential for the solution of language testing problems. When Hughes addresses the issue of content validity, he states that only if a test includes a proper sample of the relevant structure, then it would have content validity. However, a relevant structure in this case is dependent of course upon the purpose of the test and in order to judge whether a test has content validity, we need to specify the skills or structures, etc. that it is meant to cover. According to Hughes (2003), criterion-related validity relates to the degree to which results on the test agree with those provided by some independent and highly dependable assessment of the students' ability. The test is validated according to this criterion measure. Criterion-related validity is important to keep in mind, because in the schools' curriculum there are various criteria that have to be followed.

As can be seen from the literature, there are issues that need to be addressed. Both Shohamy (1995,1998) and Uysal (2010) take a critical perspective on language testing. Shohamy (1998), conducted a study in 1993 where she examined the use of an EFL (English as a Foreign Language) oral proficiency test. The test was used for graduation from secondary school, and it consisted of role play, a monologue, and an interview. An EFL inspector stated that the purpose of introducing the test was to draw teachers´ attention to oral language, a field which had been forsaken for a period of time. The impact of the test showed that the goal was achieved when it came to the fact that teachers spend more time teaching oral language. However, the teaching included only the tasks that appeared on the test, namely, role plays, monologues, and interviews. The consequence of this was that the teaching had not focused on "oral language", but more on "oral test language" and therefore became the concerning fact to oral knowledge.

In 1996 a modified version of the test was introduced. This time it involved an extensive reading component, the role play had been changed into "modified" role play where

students ask the tester questions and an extended interview instead of the interview and monologue was conducted. The purpose now as stated by the EFL inspector was: “to encourage students to read, to provide an opportunity for authentic speech and communication and to gauge the pupils' overall level of oral proficiency” (Steiner, 1995) in Shohamy (1998, p.335). The result of the study that examined the effect of this test showed that it triggered a tremendous impact on classroom activities, time allotment, and finally content and methodology. Teachers claimed that their focus was on teaching exclusively the oral skills of the exam. One of the statements was: “Of course I teach the tasks for the exam, we have no choice but to teach as dictated by the exam” (Shohamy, 1998, p.336).
However, even though that some teachers were critical about the quality of the test, they could still appreciate the status that was attached to the test. Teachers felt that the test gives oral proficiency status and did not want the Ministry to cancel the test (Shohamy, 1998).

These tests that have just been described, have been critizised by Shohamy. She states that in none of these cases were language tests used to assess language proficiency. There was no attention paid to the results in terms of language proficiency, neither students nor teachers were given any feedback or diagnosis which could have served as a formative purpose. Instead, the language tests were used as triggers which means that administrators' agendas could be conducted. The power of tests enables them to be used by bureaucratic agencies for all the described purposes (Shohamy, 1998).

Furthermore, Shohamy (1995) raises a number of questions which she considers to be important when constructing a performance language test:

> *How can the evaluation criteria reflect the kinds of judgments and consequences that the performance would entail? What relative weighting should be given to the different criteria? How can the scoring information be interpreted and presented so as to give maximum back to the test users?* (Shohamy, 1995, p. 191).

Moreover, she also discusses the questions that are more generally related to the criteria by which the performance should be judged:

> *What is the proportion of 'language' vs 'domain knowledge' to be assessed? Who should be the judge of the performance - a native speaker, a domain specialist, or a teacher?* (Shohamy, 1995, p. 191).

Even though most performance tests use the native speaker as the top level of this scale

(ACTFL, 1986, Emmett, 1985, in Shohamy, 1995) this issue has for many years been the topic in debates in the language testing literature (Alderson,1980, Bachman,1990, in Shohamy,1995). Hamilton, et.al.,1993 (in Shohamy, 1995) claim that

> *performance on a test involves factors other than straight second language proficiency, and since these factors are included in the assessment, it is expected that there will be an overlap in the performance of native and non-native speakers. Therefore the reference to native speaker performance is unwarranted* (Shohamy, 1995, p. 191).

Uysal (2010) criticize the IELTS writing test. The IELTS (International English Language Testing System) is one of the most used large-scale ESL (English as a Second Language) tests, which offers a direct writing test component. She points out the importance of drawing attention to certain issues regarding the assessment procedures of the IELTS. Uysal´s focus is especially on different reliability issues such as single marking of papers, readability of prompts, comparability of writing topics, and validity issues such as the definition of the 'international writing construct', without thinking about genres and different rhetorical conventions worldwide. Furthermore, she also discusses validity-impact issues.

Reliability issues

Even though the IELTS high stakes international writing test data reported a high reliability measure, Uysal claims that single marking is not adequate. In writing assessment it should be multiple judgements over single judgements in order to get a final score which is closer to a true score. Therefore, multiple raters should rate the IELTS writing test for inter-rater reliability (Uysal, 2010).

The IELTS pre-tests the tasks to make sure that they match the test requirements in terms of content and level of difficulty. O'Laughlin and Wigglesworth, 2003 (in Uysal, 2010) examined 'task difficulty' in Task 1 in IELTS academic writing. In terms of the language used, they found differences among tasks. They found that simpler tasks with less information developed a higher performance and more complex language from the students. On the other hand, Mickan, Slater, and Gibson, 2000 (in Uysal, 2010) investigated the 'readability of prompts' in terms of pragmatic and discourse features and the test takers 'test-taking behaviours' in the writing test. They found that the two things which influenced the writing performance and task comprehension were the purpose and the lexicogrammatical structure.

Mickan (2003) (in Uysal, 2010) addressed the issue of inconsistency in ratings in IELTS exams. He spotted difficulties in identifying certain specific lexiogrammatical features that specifies various levels of performance. Furthermore, he discovered that even though there was an analytical scale used, raters had a tendency to respond to texts more as a whole rather than looking at all the individual components. Noteworthy is that the IELTS claims that "the use of analyctic scales contributes to higher reliability as impressionistic rating and norm referencing are discouraged, and greater discrimination across bands is achieved" (Uysal, 2010, p.316)

Validity issues

The IELTS claims that it is an *international* English test. The claims are based on the following issues:

1. *Reflecting social and regional language variations in test input in terms of content and linguistic features, such as including various accents.*
2. *Incorporating an international team (UK, Australia, and New Zealand) which is familiar with the features of different varieties in the test development process.*
3. *Including NNS (Non Native Speakers) as well as NS (Native Speakers) raters as examiners of oral and written tests* (Uysal, 2010, p.317).

However, according to Taylor, 2002, (in Uysal, 2010) the construct definition does not vary from other language tests. If IELTS claims that the purpose is to assess international English, then evidence is needed to support that claim and moreover include international language features in the construct definition. Furthermore, Taylor suggests that discourse variations may occur across cultures. As a result the IELTS writing test should think about the differences in rhetorical conventions and genres around the world. A genre is not universal, but culture specific. The argument styles, logical reasoning, organizational patterns, rhetorical norms, etc. varies in different parts of the world.

Investigations have been made regarding the consequences and impact of the content and nature of classroom activity in IELTS classes. Moreover, test takers and test users attitudes have been examined. Uysal, however, states that there is a lack of investigation and that the impact of writing tests when speaking of chosen standards or criteria on the international communities should be given more attention (Uysal, 2010).

With this said, it is quite clear that high-stakes tests like the EFL oral proficiency test

and IELTS writing test has its drawbacks and should be approached with a critical point of view according to Shohamy (1995, 1998) and Uysal (2010)

5.1.3 Evidence in validity

Weir (2005) states that the satisfactory evidence of validity is highly necessary for any serious test. Then he addresses two different concepts when he describes the importance of validity. Concept 1. *Validity resides in test scores*. By this he means that validity in this case might be better defined as the extent to which a test could produce proper data, i.e., test scores which are accurate in their representation of what level a student is at regarding their skills or knowledge of the language. His point is that it is improper to discuss whether tests like Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS) are valid or not. He is more concerned about the scores produced by a particular administration of a test on a particular sample of candidates. Then over time, cases can be made that different tests are valid if various versions of a test or administrations of a test show similar results.

Concept 2. *Validity is multifaceted*. Weir explains this concept by saying that to support any claims for the validity of scores on a test, there is a need of different types of evidence. As an evidential basis for test interpretation, these are complimentary aspects and not alternatives. One single validity aspect may not be looked upon as better or superior to another. If there is deficit in any one, then it might raise questions regarding how well-founded the interpretation of test scores are.

As we can see from the presented research above, one might think that when we use the term validity in testing, it might seem as if we should consider validity as a checklist procedure. However, that is not the case. As Haertel (in Lane,1999) points out, when accumulating validity evidence, it should be treated as more than a checklist procedure. Haertel, Messick, Cronbach, Kane, Crooks, and Cohen (in Lane,1999) point out that the validation process involves the development and evaluation of a coherent validity argument for and against proposed test score interpretations and uses. In the validity argument, each inference is based on a proposition or assumption that requires support. When we set forth a validity argument, it allows for the accumulation of evidence not only for but also against test score interpretations. Messick (in Lane,1999) states that the process of validation involves gathering evidence for and looking into possible threats to the validity of test score interpretations. Furthermore, Kane, Crooks, and Cohen (ibid) argue that "...the most attention

should be given to the weakest part of the interpretative argument because the overall argument is only as strong as its weakest link" (Lane, 1999, p.1). Moreover, Lane argues that to establish what validity evidence is necessary for a particular purpose of testing, analysts should define a set of propositions that would support the proposed interpretation. For each proposition evidence should then be collected as support. As an example Lane uses a high school certification test. This test was developed to determine if students mastered the state content standards. She lists four relevant propositions:

1. *The test content is representative of the state content standards.*
2. *The test scores can generalize to other relevant set of items.*
3. *The test scores are not unduly high or low due to irrelevant constructs being measured.*
4. *The students curriculum and instruction afforded them the opportunity to attain the state content standards* (Lane, 1999, p.2)

As we can see from this example, different sources of evidence can be accumulated and used to either verify or deny a validity argument. To determine to what extent a validity argument is supported, it is important that the evidence is not collected in a gradual fashion, but should be evaluated continuously.

5.2 Reliability

In the previous chapter the concept of validity has been discussed. This following chapter will focus on the term reliability. It is often argued that the two concepts are compatible since a test needs to be reliable in order to be valid, even though the reverse is not necessarily true. Davies (in Alderson & Banerjee,2002) argues that if reliability is maximized it may be at the expense of validity, and if validity is maximized then it might be at the expense of reliability when it comes to language testing. There is a distinction between reliability and validity and Alderson problematises this. He claims that even though in theory the difference is clear, problems occur when we consider how reliability is measured. Swain (as cited by Alderson & Banerjee, 2002) also argues that "since SLA research establishes that inter-language is variable, the notion of internal consistency as a desirable feature of language tests is highly questionable" (Alderson & Banerjee, 2002, p.101) By this we can draw the conclusion that high internal consistency indicates low validity in her opinion.

It is said that even though test-retest reliability is the easiest way to measure

reliability, there are problems with that concept. For example, if a student takes the same test a second time and the test is reliable then the score should remain constant. But what if the score changed because the student learned from the first administration or because the student´s ability had somehow changed. In any case, we might expect a somewhat lower test- retest relation. That would be considered as a valid indication of the change in ability. According to Alderson (ibid) is not clear that this example represents a lack of reliability. Another way to measure reliability is to use parallel forms of the test. However, parallel forms of tests are often *validated* by correlations (concurrent validity) so therefore high correlations between parallel forms are more a measure of validity and not so much reliability (Alderson & Banerjee, 2002).

5.2.1 Different types of reliability

When Strong (1995) describes reliability, he discusses five different basic types where several of them can be tested statistically. The first is the *test-retest* where he uses the term hypothetical question about the degree of correlation between test scores if a student took the same test twice. Due to chance or maybe error in administrating the text, there will be variations in scores in every test. So, the degree of reliability in a test's administration would be the correlation between the two different scores. However, this has its drawbacks considering the fact that students would probably remember many of the questions from the first test they took. As we can see, the conclusion Strong draws regarding drawbacks when it comes to test-retesting is similar to the one that Alderson and Banerjee made. The second type of reliability Strong mentions is *parallel tests* (also discussed by Alderson and Banerjee). Strongs definition here is the extent to which any two forms of the same test measure the same skills or traits. The third is the *internal consistency* of a test, which means to what extent test questions measure the same skills or traits and that the test questions are related to one another. The last two kinds of reliability on a test are "*inter-rater reliability*" and "*intra-rater reliability"* These two types refer to the scoring done on subjective tests. Inter-rated reliability is the degree to which two different markers or raters agree on a score for a student paper. Intra-rated reliability on the other hand is when one rater or marker scores consistently from one student's paper to another.

In terms of these last two types of reliability, Strong claims that there is overwhelming evidence that the scoring of writing as an example is highly unreliable unless certain procedures are followed.

1. *Setting the scoring criteria in advance.*
2. *Providing sample answers for the markers*
3. *Training the markers to use the criteria*
4. *Scoring each paper twice, and a third time if the contrast is too big in the scores attributed to the same paper* (Strong, 1995, p.9).

Uysal (2010) criticize the IELTS (International English Language Testing System) writing test. Her focus is on reliability and validity issues. She claims that multiple raters and judgements are to prefer over single raters and judgement.

5.3 Washback

Spolsky (in Bailey, 1999) claims that it has been argued for many years, within a broad context, that tests have a powerful influence on language learners who are preparing for exams, and the teachers who are helping them in their preparation. These three following statements are typical claims which can be found in the most accounts:

> *It is generally accepted that public examinations influence the attitudes, behavior, and motivation of teachers, learners, and parents... (Pearson in Bailey, 1999, p.1)*

> *It is common to claim the existence of washback (the impact of a test on teaching) and to declare that tests can be powerful determiners, both positively and negatively, of what happens in classrooms. (Wall & Alderson in Bailey, 1999, p.1)*

> *The washback effects of large-scale testing programs on instruction are widely discussed. In the view of instructors and students, such tests contain what students must learn and therefore what must be taught- a reasonable view, given that the tests in many cases represent the language hurdle students must clear before continuing their academic careers. (Chapelle & Douglas in Bailey, 1999, p.1)*

The definitions of washback and related concepts are almost as many as the people who write about it. Some definitions are very simple and easy to understand, while other definitions are very complex. Some tend to focus on teachers and students in the classroom environment,

while other, more complex, tend to involve references to what influences tests might have on the educational system or even society in general (Bailey,1999). Alderson and Banerjee (2001) approach washback more straightforward. Here is their definition of washback:

The term ´washback´ refers to the impact that tests have on teaching and learning. Such impact is usually seen as being negative: tests are said to force teachers to do things they do not necessarily wish to do. However, some have argued that tests are potentially also ´levers for change´ in language education: the argument being that if a bad test has negative impact, a good test should or could have positive washback (Alderson & Banerjee, 2001, p.214)

As previously mentioned, there are many different views on washback. Alderson and Banerjees (2001) view was quite straightforward whereas Shohamys (in Bailey, 1999) take on washback could be considered as slightly more complex. Shohamy has summarized four key definitions:

1. *Washback effect refers to the impact that tests have on teaching and learning.*
2. *Measurement driven instruction refers to the notion that tests should drive learning.*
3. *Curriculum alignment focuses on the connection between testing and the teaching syllabus.*
4. *Systemic validity implies the integration of tests into the educational system and the need to demonstrate that the introduction of a new test can improve learning (Bailey, 1999, p.3).*

In another article, Shohamy (in Bailey,1999) contrasts school tests and external tests. She notes that

> *external tests have become most powerful devices, capable of changing and prescribing the behaviour of those affected by their results-administrators, teachers and students. Central agencies and decision makers, aware of the authoritative power of external tests, have often used them to impose new curricula, textbooks, and teaching methods. Thus external tests are currently used to motivate students to study, teachers to teach, and principals to modify the curriculum. The use of external tests as a device for affecting the educational*

process is often referred to as the washback effect or measurement- driven instruction. (Bailey, 1999, p.4).

Bachman and Palmer (1996) notes that even though most discussions regarding washback has focused on processes (learning and instruction), their perspective is that washback can be considered best within the scope of *impact.* They discuss the impact on individuals. In this sense the focus is on those individuals who are most directly affected by test use: test takers and teachers. Moreover, they also discuss the impact on society and education systems.

5.3.1 Impact on test takers

According to Bachman and Palmer (1996), the testing procedures can be viewed from three different aspects where test takers can be affected:

1. *The experience of taking and, in some cases, of preparing for the test,*
2. *The feedback they receive about their performance on the test, and*
3. *The decisions that may be made about them on the basis of their test scores (Bachman &Palmer, 1996, p.31).*

Bachman and Palmer explain the first aspect by using high-stakes tests such as standardized tests and public examinations as an example. In these tests the test takers might spend several weeks preparing for the test individually. In several countries where high-stakes tests might be used as a selection for higher levels of school or placement into universities, the teaching may have its focus on the syllabus of the test many years before the actual test takes place. The techniques that are required in the test will be practiced in class.

5.3.2 Impact on teachers

The second aspect of individuals who are directly affected by tests are the test users which in this case concern the teachers. Here Bachman and Palmer (1996) discusses the term "teaching to the test" which means teaching that is not compatible with teachers´ own values or goals, or with the values and goals of the instructional program. By looking at it from this perspective, if teachers feel that what they teach is not relevant to the test (or the other way round) then the test might lack in authenticity. In that case, the washback of the test may be harmful, or have a negative impact on instruction. There might also be dissatisfaction based on test results or because of the fact that various aspects of the program such as: curriculum, materials, types of

learning activities etc. might not be in line with what teachers believe promotes effective learning. This dissatisfaction arises from those who are responsible for the instructional program. When it comes to situations like this, a number of language testers argues that a way to bring instructional practice compatible with current thinking in the field is to develop a testing procedure that reflects this thinking.

Moreover, Shohamy (1998) studied the EFL (English as a Foreign Language) oral proficiency test and stated that the teachers were critical towards the test. They felt as if they were only "teaching for the test" considering the fact that the specific tasks that appeared on the test were only practiced. With this background Shohamy claimed that this language test was not used to assess language proficiency.

5.3.3 Impact on society and education system

The societal and educational value systems that inform the test use must always be considered by test users and test developers. The values and goals becomes very complex in the context of second or foreign language testing, since the values and goals that inform test use varies from different cultures. For example, one culture may emphasize individual effort and achievement, while another culture might emphasize group cooperation and respect for authority. Another aspect that needs to be considered is the consequences of our actions. We must realize that when we use a language test, it is likely to have consequences not just for the individual, but also for the educational system and society. This is of great significance when it comes to high-stakes testing (Bachman, Palmer, 1996).

In addition to this, McNamara (2000), discusses that the power of tests influences the reputation of teachers and schools, which could lead to a strong influence on the curriculum. McNamara states that ethical language testing should work to ensure positive washback from tests. However, sometimes the responsible authorities use assessment reform to drive curricular reform, believing that assessment can be designed to have positive washback on the curriculum.

6 Conclusion and discussion

As McNamara (2000) stated, testing is a universal feature of social life and testing for special purposes or to establish identity has become an accepted part of many fields. What is true of

testing in general is true also in language testing. The purpose of this literature review was to examine what different important aspect we need to consider when constructing a language test based on research presented in various literature. In the research reviewed, I discovered that the key elements validity, reliability, and washback are highly relevant when constructing and using a language test in order for the test to have the proper effect on both test takers and test users. However, although there are many aspects of these concepts most studies do not differ in a large scale.

It is important to keep in mind that even though both high-stakes tests and classroom testing includes these concepts, they are sometimes objects for criticism and questioning. As we can see from the presented research by both Shohamy and Uysal, there are issues that has to be addressed. To exemplify this, Shohamy (1995) raised a number of questions which she considers to be important when it comes to constructing a language test. For example, how scoring information can be interpreted and presented so the test user can receive the ultimate feedback, and whether a native speaker should be the judge of the performance or the teacher. Another issue that is mentioned by Uysal (2010) is that the IELTS (International English Language Testing System) writing test should consider the differences in rhetorical conventions and genres around the world. A genre is not universal, but culture specific.

After reviewing the literature regarding the concepts of validity, reliability, and washback, these are conclusions that can be drawn: First, it is clear that when the term *validity* is discussed, we need to keep in mind that when tests are used, they have to measure what they were supposed to measure. Otherwise the test could be an object for questioning and highly criticized not just among test takers and test users, but also from higher instances. Furthermore, the *evidence in validity* is of great significance. Weir (2005) stated that the provision of satisfactory evidence of validity is indisputably necessary for any serious test and Lane (1999) also argued for the importance of the need for evidence in validity. Hughes (2003) takes this further and states that it is not sufficient enough to just claim that a test has validity, but we also need empirical evidence that strengthens the validity claim.
Secondly, there have been discussions whether *reliability* and *validity* should be treated as one joining factor. However, in the literature that has been examined those two factors have been addressed separately. When the literature regarding reliability was examined, the terms *test-retest* and *parallel tests* were emphasized. These two ways to measure reliability are considered to be fairly easy ways of measurement. But despite this, there are problems that

need attention. Alderson and Banerjee (2002) stated that: what conclusions can we draw if a student takes a test twice, and the result varies because the ability of the student had changed in some way, or that the student had learned from the first test?

Finally, when the research on *washback* was investigated it is quite clear that it has a powerful influence on test takers, test users, and society and educational systems and it is suggested that the focus should be on *impact* and not on *processes* as Bachman and Palmer (1996) stated. In this regard the individuals who are most affected by test use test taker and teachers. The problems that could occur here is when the teachers "teach to the test". The consequence of this could be that the authenticity of the test becomes a matter of questioning and therefore the washback could have a negative effect.

With this said, we can draw parallels to the relevance this has in Swedish schools. For example, the national exams in English are a major part when it comes to the grading of the students.

Not only is this a matter of importance when taking high-stakes tests like the national exam, but also when it comes to the various tests that are conducted in the classroom. Today's students might ask the questions why do we need this?, what are we supposed to do?, or how are we supposed to do it? In that case, the teacher has to be able to justify the choices by pointing at the different criteria and feel comfortable knowing that the foundation of the test is valid and reliable. However, even though the teacher might feel satisfied with the choices made, we can rest assure that most classroom tests are neither hundred per cent valid nor reliable because teachers are not able to construct tests that have all these features. So, the question is how the teachers can be better prepared for this issue. According to my review on the literature, there is a clear gap of research regarding the matter of classroom testing and how the teachers prepare tests. The national exams are considered to be well prepared (even though they also have problems). However, classroom testing and how to increase the teachers' awareness when constructing a language test is something that deserves more attention. I would suggest that future studies put more focus on the teachers' aspect and point of view since I consider this to be a major factor in my own future profession as a teacher. Moreover, we need to address the issue of "teaching to the test". Teachers should promote "life-long learning" and try to increase students' motivation by stressing that the tasks and assignments that are held in class, does not have the purpose of only passing a certain test, which many times seems to be the only reason for students.

Reference list

Alderson, C. J., & Banerjee, J. (2001). *Language testing and assessment (Part 1). Language Teaching*, 34, pp 213-236 doi: 10.1017/S0261444800014464

Alderson, C.J., & Banerjee, J. (2002). *Language testing and assessment (Part 2). Language Teaching*, 35, pp 79-113 doi: 10.1017/S0261444802001751

Bachman, L. F., & Palmer, A. S. (1996).*Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bailey, K, M. (1999). *Washback in language testing.* Research Monograph 99-4, Princeton N.J.: Educational Testing Service.

Dragemark Oscarson, A. (2009). *Self-assessment of writing in learning english as a foreign language: A study at the upper secondary school level.* Acta Universitatis Gothoburgensis.*277*. Doctoral thesis.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Korp, H., & Sverige. Myndigheten för skolutveckling. (2003).*Kunskapsbedömning: Hur, vad och varför / [elektronisk resurs]*. Stockholm: Myndigheten för skolutveckling.

Lane, S. (1999). *Validity evidence for assessment*. Reidy Interactive Lecture Series. McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Shohamy, E. (1998). Critical language testing and beyond.*Studies in Educational Evaluation,24*(4), 331-345. doi:10.1016/S0191-491X(98)00020-0

Shohamy, E. (1995). Performance assessment in language testing.*Annual Review of Applied Linguistics,15*, 188-211. doi:10.1017/S0267190500002683

Strong, G. (1995). *A survey of Issues and Item Writing in Language Testing.* Retrieved 2014- 10-15, from http://eric.ed.gov/?id=ED397635

Uysal, H. H. (2010). *A critical review of the IELTS writing test*. *ELT Journal,64*(3), 314-320. doi:10.1093/elt/ccp026

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.